Forecasting Stock Price Using Machine Learning Technique

Mohammed Nazim Uddin, PhD
School of Science, Engineering and Technology
East Delta University
Chattagram, Bangladesh
nazim@eastdelta.edu.bd

Md Tanvir Rahman
School of Science, Engineering and Technology
East Delta University
Chattagram , Bangladesh
Ornobtanvir.git@gmail.com

Abstract*—Stock market is an emerging sector in any country of the world . Many people directly related to this sector . Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument .When publicly traded companies issue shares of stock to investors, each of those shares is assigned a monetary value, or price. Stock prices can go up or down depending on different factors. Stock prices can be affected by a number of things including volatility in the market, current economic conditions, and popularity of the company. The successful prediction of a stock's future price could yield significant profit .Along with the development with the stock market ,forecasting become an important topic .Since finance market has become more and more competitive, stock price prediction has been a hot research topic in the past few decades .predicting stock price is regarded a challenging task because stock market is essentially non linear ,non-parametric,noisy,and a chaotic system .Trend of a market depends on many things like liquid money human behavior, news related to stock market etc. All this together controls the behavior of trends in a stock market with the advancement of the computing technology we use machine learning technique,like Support Vector Regression,K-nearest-neighbor,liner Regression, Random forest Regression , for analyzing time series data to predict stock price. In this paper we try to develop a forecasting model with stacking multiple method to find the best forecast of the stock price.

***Keywords*— Time Series data, SVR, KNN-Regressor, liner regression,staking regressor,Random forest regression'**

I. Introduction

The goal is to take time series data, find the equation that best fits the data, and be able forecast out a specific value. Time series data is a continuous data statistical observations recorded over a specific period of time. This model will try to understand the pattern of the continuous data[1] by combining different method and produce a best fit line that fits the data. The target is to determine the future stock price and improve their strategy for future. regression[2] models are among the most known regression models used in the machine learning community and recently many researchers have examined their sufficiency in ensembles[3][4]. Although many methods of ensemble design have been proposed, there is as yet no obvious picture of which method is best. One notable successful adoption of ensemble learning is the distributed scenario. In this work, we propose an efficient distributed method that uses the same training set with the parallel usage of an averaging methodology that combines linear regression[5] and KNN regression models[6],Support Vector Regression[7],random Forest Regresssion[8][9]. We performed a comparison of the presented ensemble with other ensembles that use either the linear regression as base learner and the performance of the proposed method was better in most cases. Using averaging methodology, we expect to obtain better results because both theory and experiments show that averaging helps most if the errors in the individual regression models are not positively correlated. linear regression is a linear approach to modelling the relationship between a scalar response dependent variable and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression support vector machines are supervised learning models algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model by creating a hyper plane[9] that assigns new examples to one category or the other . in support vector regression (SVR). The model produced by support vector classification (as described above) depends only on a subset of the training[10] data, because the cost function for building the model does not care about training points that lie beyond the margin. In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether KNN is used for classification or regression.In case of the knn regression the output is the property value for the object. This value is the average of the values of its k nearest neighbors.a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones A model that combines KNN regression,Linear regression,Support Vector Regression, Random Forest regression model used for predicting stock prices can forecast better price accuracy.The paper is structured as follows. Section II presents related work on the Stock price prediction. Section III presents the most well-known methods for building ensembles .Section IV contains Proposed methodology. Section V contains Architecture Overview,Section VI contains Implementation and Evaluation . Section VII contains conclusion

II .RELATED WORK

A lot of research has been done to predict the stock price. Many algorithms of data mining have been proposed to predict stock price. Neural Network, Genetic Algorithm, Decision Tree and Fuzzy systems are widely used. pattern discovery is beneficial for stock market prediction and public sentiment is also related to predicting stock price. There is a certain correlation among them. previous studies on stock price forecasting shows the prevalent use of technical indicators with artificial neural networks (ANN) for stock market prediction One of the well-researched and most important algorithm in the field of Data mining is Association Rule Mining (ARM), Decision trees are excellent for making financial decisions where a lots of complex data needs to be taken into account. They provide an effective framework in which alternative decisions and the implications of taking those decisions can be laid down and evaluated. They also form an accurate, an algorithm called AIS was proposed for mining association rules [11]. For last fifteen years many algorithms for rule mining have been proposed. Wanzhong Yang [12] also proposed one innovative technique to process the stock data named Granule mining technique, which reduces the width of the transaction data and generates the association rules. R.V.Argiddi [13] has proposed fragment based mining which deals mainly with reducing the time and space complexity involved in processing the data in association rule mining technique. As in granule mining, fragment based approach fragments the data sets into fragments for processing thereby reducing the input size of data sets fed to the algorithm. In contrast to granule mining, in fragment based mining the condition and decision attributes are summed for obtaining generalized association rules. Kannika Nirai Vaani M,E Ramaraj [14] has now proposed new approach to generate association rules Providing faster generation of frequent item sets to offer interesting and useful rules in an effective and optimized way with the help of Genetic Algorithm approach. From the above literature review, technical indicators with different data mining techniques had been widely used, while there are only few studies of the use of fundamental indicators. The impact of fundamental analysis variables has been largely ignored like Price earnings ratio, Moving average, rumors etc. But Prashant S. Chavan , Prof. Dr. Shrishail. T. Patil [15] has said that hybridized parameters gives better & more accurate results that applying only single type of input variables. A.A. Adebiyi , C.K. Ayo, M.O Adebiyi and S.O. Otokiti [16] has proposed predictive model has the potential to enhance the quality of decision making of investors in the stock market by offering more accurate stock prediction using hybrid parameters. They used ANN for this but their performance is not always satisfactory. Robert K. Laia , Chin-Yuan Fanb, Wei-Hsiu Huang b, Pei-Chann Chang [6,17],[9,18] has proposed forecasting model that integrates a data clustering technique, fuzzy decision tree (FDT) and genetic algorithm (GA) to construct a decision-making system based on historical data and technical indices.Public information such as news, blogs, twitter mood, social networking sites and stock articles can also affect stock market trend. Web has been treated as a great source of financial information; many papers proposed stock price predicting approaches based on analyzing web sentiments using text mining.Schumaker and Chen examined different textual representations of news articles to predict future stock price, which was compared to linear regression with SVM.

.

III .Ensembles of Regression Models

Bagging is a "bootstrap" ensemble method that creates indivrated by randomly drawing, with replacement, N examples - where N is the size of the original set; many of the original examples may be repeated in the resulting training set while others may be left out. After construction of several regression models, averaging the predictions of each regression model performs the final prediction . Instability (responsiveness to changes in the training data) is a prerequisite for bagging to be effective Another approach for building ensembles of regression models is to use a variety of learning algorithms on all of the training data and combine their predictions. When multiple regression models are combined using averaging methodology, we expect to obtain good results based on the belief that the majority of experts are more likely to be correct in their decision when they are close in their opinions Stacked generalization or Stacking, is a more sophisticated approach for combining predictions of different learning algorithms. Stacking combines multiple regression models to induce a higher-level regression model with improved performance. In detail, the original data set constitutes the level zero data and all the base regression models run at this level. The level one data are the outputs of the base regression models. A learning algorithm is then used to determine how the iduals for its ensemble by training each regression model on a random redistribution of the training set. Each regression model's training set is generated by randomly drawing, with replacement, N examples - where N is the size of the original set; many of the accurately learned. After several cycles, the prediction is performed by taking a weighted average of the predictions of each regression model, with the weights being proportional to each regression model's performance on its training set.

IV. PROPOSED METHODOLOGY

We generate linear regression algorithm from this formula

$$m=(N\Sigma xy - \Sigma x\,\Sigma y)\,/N(\Sigma x^2) - (\Sigma x)^2 \quad (1)$$

$$b = \Sigma y - m(\Sigma x)\,/N \quad (2)$$

$$y = mx + b \quad (3)$$

Then we generate knn-regression by computing euclidean distance from the query example to the labeled example. and ordering the example by increasing rate .we estimate euclidean distance by following formula

$$dist(A,B)=\sqrt{\sum(xi-yi)/2m} \quad (4)$$

Then generate SVM regression, the input (X)tis first mapped onto a *m*-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is constructed in this feature space. Using f(x,w) mathematical notation, the linear model (in the feature space) is given by where g(x) denotes a set of nonlinear transformations, and *b* is the “bias” term.

$$f(\mathbf{x},\omega)=\sum_{j=1}^{m}\omega_j g_j(\mathbf{x})+b$$

loss function is estimated by this formula

$$L_{\varepsilon}(y, f(\mathbf{x}, \omega)) = \begin{cases} 0 & \text{if } |y - f(\mathbf{x}, \omega)| \leq \varepsilon \\ |y - f(\mathbf{x}, \omega)| - \varepsilon & \text{otherwise} \end{cases}$$

Then stacking is performed using these algorithm. Stacking is concerned with combining multiple classifiers generated by different learning algorithms $L_1, \ldots. L_n$ on a single data set S, which is composed by a feature vector $S^i = (X^i, Y^i)$.

• The stacking process can be broken into two phases:

1. Generate a set of base-level classifiers $C_1, \ldots, C_n$.

• Where $C_l = L_l(S)$

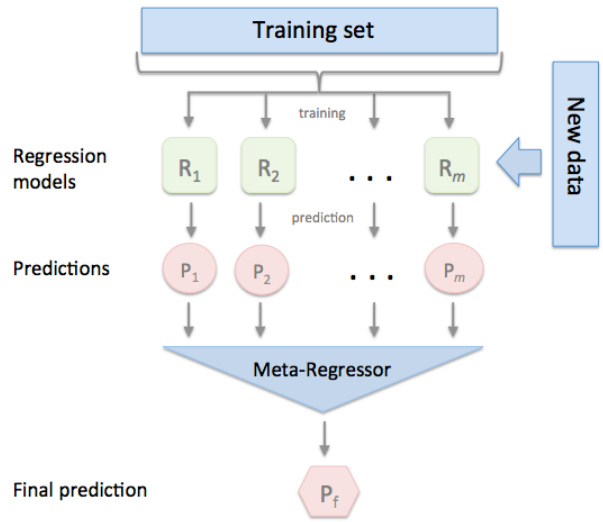
2. Train a meta-level classifier to combine base level classifier

3.The training set for the meta-level classifier is generated through a leave-one-out cross validation process.

$$\forall^i = 1, \ldots n \text{ and } \forall_k = 1, \ldots, N$$

$$C^i_k = L_k(S - s^i)$$

• The learned classifiers are then used to generate predictions for $Y_k^i = C^i_k(x^i)$



• The meta-level data sets consists of examples of the form where the features are the predictions of the base-level classifiers and the class is the correct class of the example in hand.

V. ARCHITECTURE OVERVIEW

The prediction system has a two tier architecture top tier is dedicated to preparing the data sets from multiple information sources to make them ready for the predication tasks in the next tier. It is composed of two major parts. The first part is data prepossessing . In this process we process the data by adding more feature and removing unnecessary feature and removing the bad data and also the absence of the data . The second part is the data alignment. The second tier is dedicated to the market volatility analysis and prediction through the model integration and training, which uses multiple kernel learning methodology to train the model It consists of three tasks: First, we build one regression model per source. Second, we train the model with the same data sets ,then we create a stacked algorithm using these algorithm. In this paper, we use the multi-kernel learning method .

A. Data Prepossessing

In the real world, many data sets are very messy.Most stock price/volume data is pretty clean, rarely with missing data, but many data sets will have a lot of missing data. filter out other unimportant feature from the feature because not all the feature will be included into the final feature list .The reason behind it is the unnecessary feature and those value which has no relation with the stock market prediction will reduce the accuracy of the prediction.

B. Feature Extraction

We only work with the valuable feature , created some new valuable feature through manipulation. Right feature can help to get more accurate results.

C. Training The ML Model

The way this works is taking 75% of data, and use this to train the machine learning classifier. Then taking the remaining 25% of our data, and test the classifier. Since this is sample data, we should have the features and known labels. Thus, if we test on the last 25% of our data, we can get a sort of accuracy and reliability

D. Creating Stacked algorithm

Create the stacked algorithm using the model previously made. And averaging the prediction for creating the final prediction.

VI. IMPLEMENTATION AND EVALUATION

A. Data Collection

Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion. Which enables us to evaluate outcome .In this case for regression we used stock price data set of Google, Apple, Microsoft, IBM, Nike which is collected from Quandl which is a platform for financial data.

VII. RESULT ANALYSIS

TABLE I. ACCURACIES OF THE PROPOSED MODEL

Company	**Accuracy Based on Cross Validation**				
	Linear regressionm	***Knn regression***	***SVR***	***RFG***	***Proposed model***
Google	87.97%	89.18%	74.32%	88.67%	90.72%
Apple	92.2%	91.4%	91.0%	93.4%	94.8%
Microsoft	91.4%	91.4%	91.4%	91.4%	91.4%
IBM	73.6%	71.3%	77.8%	78.4%	80.8%
Nike	94.08%	94.9%	95.05%	92.9%	96.8%

After determining and comparing with other models. In our proposed model we have attained the highest accuracy among all others.

VIII. CONCLUSIONS AND FUTURE WORK

It is known that if we are only concerned for the best possible correlation coefficient, it might be difficult or impossible to find a single regression model that performs as well as a good ensemble of regression models. In this study, we built an ensemble of regression models using four different learning methods.

REFERENCES

[1] Classification with discrete and continuous variables via general mixed-data models ,Alexander de Leon:University of Calgary,Andrea Soo: University of Calgary

[2] J. Fox, Applied Regression Analysis, Linear Models, and Related Methods, ISBN: 080394540X, Sage Pubns (1997).

[3] L. Breiman, Bagging Predictors. Machine Learning, 24(3) (1996) 123-140.

[4] D. Opitz, R. Maclin, Popular Ensemble Methods: An Empirical Study, Artificial Intelligence Research, 11 (1999): 169-198, Morgan Kaufmann. 12.C. Perhch, F. J. Provost, J. S.

[5] Linear regression Analysis on Net Income of an Agronomical Company ,Supichaya Sunthornjittanon:Portland State University

[6] Efficient and Accurate knn based Classification and Regression,Harshit Dubey:International Institute of Information Technology

[7] Design and Training of Support Vector Machines,Alistair Shilton:The University of Melbourne

[8] Advances in Random Forests with Application to Classification,Arnu Pretorius:Stellenbosch University

[9] Cristianini and J. Shawe-Taylor.An Introduction to Support Vector Machines and other kernel-based learning method. Cambridge University Press, Cambridge, UK, 2000

[10] L. Breiman, Stacked Regression. Machine Learning, 24 (1996):49-64.4.T.G. Dietterich, Ensemble methods in machine learning. In Kittler, J., Roli, F., eds.:

[11] R. Agrawal, T. Imielinski, and A.Swami (1993) “Mining association rules between sets of items in

large databases”.

(ACM SIGMOD '93).

[12] Wanzhong Yang. “Granule Based Knowledge Representation for Intra and Inter Transaction Association Mining”, Queensland University of Technology, July 2009.

[13] R.V.Argiddi , S.S.Apte “a study of association rule mining in fragmented item-sets for prediction of transactions outcome in stock trading systems” IJCET-2012.

[14] KANNIKA NIRAI VAANI M, E RAMARAJ “ AN ENHANCED APPROACH TO DERIVE DISJUNCTIVE RULES FROM ASSOCIATION RULE MINING WITHOUT CANDIDATE ITEM GENERATION” IJCA-2013.

[15] Prashant S. Chavan, Prof. Dr. Shrishail. T. Patil” Parameters for Stock Market Prediction” IJCTA | (Mar-Apr 2013).

[16] Multiple Classifier Systems. LNCS Vol. 1857, Springer (2001) 1-1 N. Duffy, D. Helmbold, Boosting Methods for Regression, Machine Learning, 47, (2002)

A.A. Adebiyi, 1C.K. Ayo, 1M.O Adebiyi, and 2S.O. Otokiti “An Improved Stock Price Prediction using Hybrid Market Indicators” ICT. (2002)

[17] J. Friedman, Stochastic Gradient Boosting,Computational Statistics and Data Analysis

[18] Y. Grandvalet, Bagging Equalizes Influence, Machine Learning, Volume 55(3) (2004)

[19] N.L. Hjort, G. Claeskens, Frequentist Model Average Estimators, Journal of the American Statistical Association,

[20] Y. Morimoto, H. Ishii, S. Morishita, Efficient Construction of Regression Trees with Range and Region Splitting, Machine Learning

[21]

[22] Simonoff,Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. Journal of MachineLearning Research 4 (2003) 211-255

[23] Y. Wang, I. H. Witten, Induction of model trees for predicting continuous classes, In Proc.of the Poster Papers of the European Conference on ML, (1997) 128-137.